

PECAN: Library Free Peptide Detection for Data-Independent Acquisition Tandem Mass Spectrometry Data

Supplementary Materials

Ying S. Ting¹, Jarrett D. Egertson¹, James G. Bollinger¹, Brian C. Searle¹, Samuel H. Payne²,
William Stafford Noble^{1,3}, and Michael J. MacCoss¹

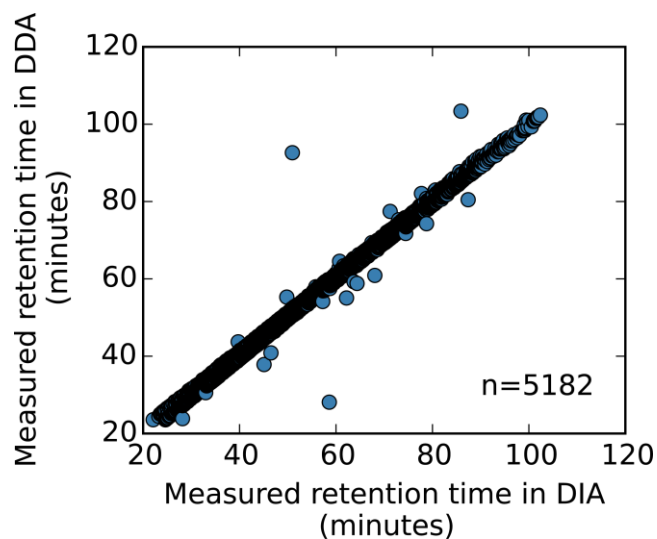
¹Department of Genome Sciences, University of Washington, Seattle, Washington, USA

²Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA

³Department of Computer Science and Engineering, University of Washington, Seattle,
Washington, USA

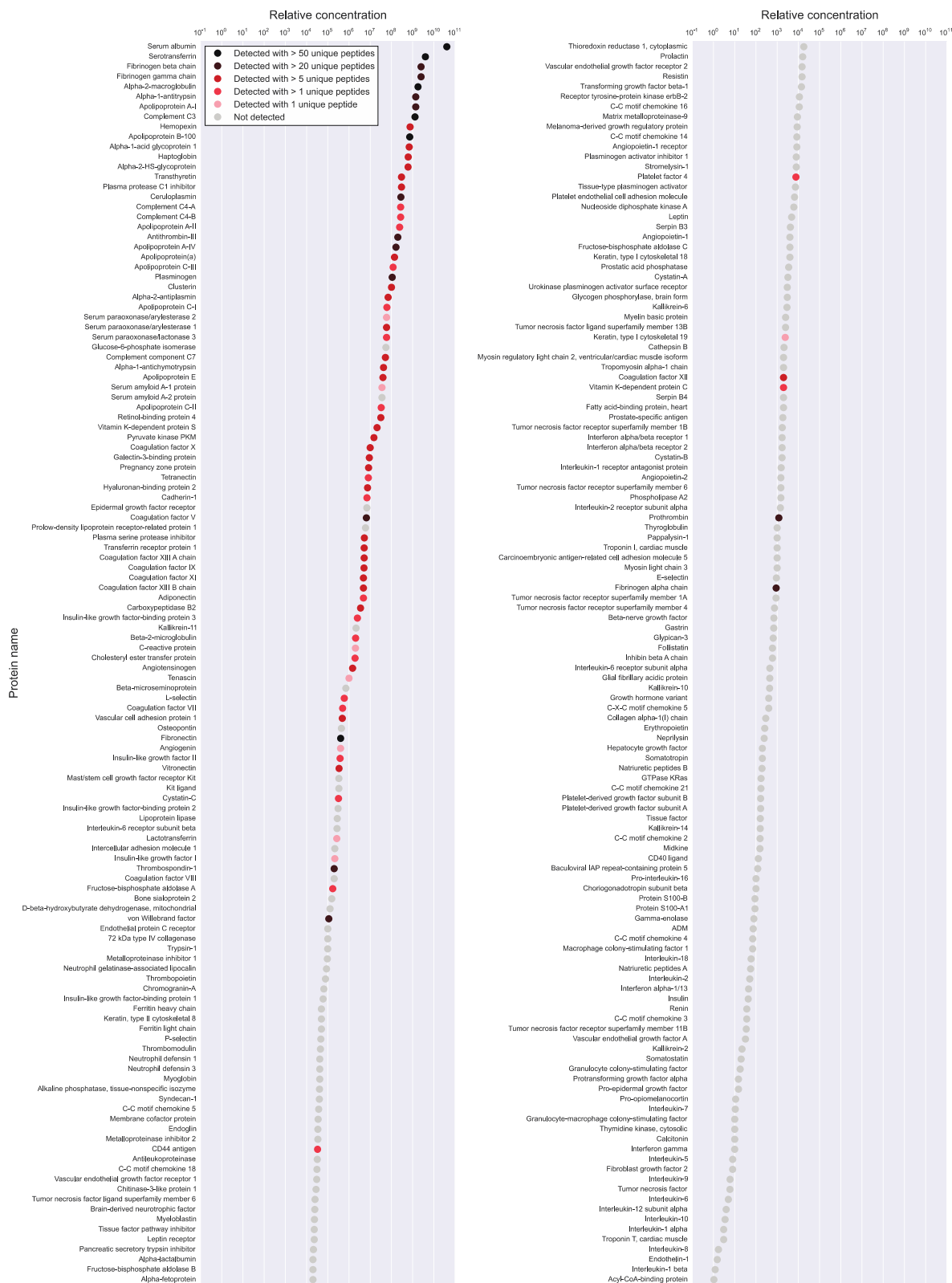
Supplementary Figures

Supplementary Figure 1 – Retention time analysis for common peptides from Comet-DDA and PECAN-DIA



Of the 5,182 peptides commonly detected by PECAN from 4xGPF DIA data and Comet from 4xGPF DDA data (Fig. 3a), 27 peptides were identified more than 2 minutes apart.

Supplementary Figure 2 – Dynamic range of DIA plasma library

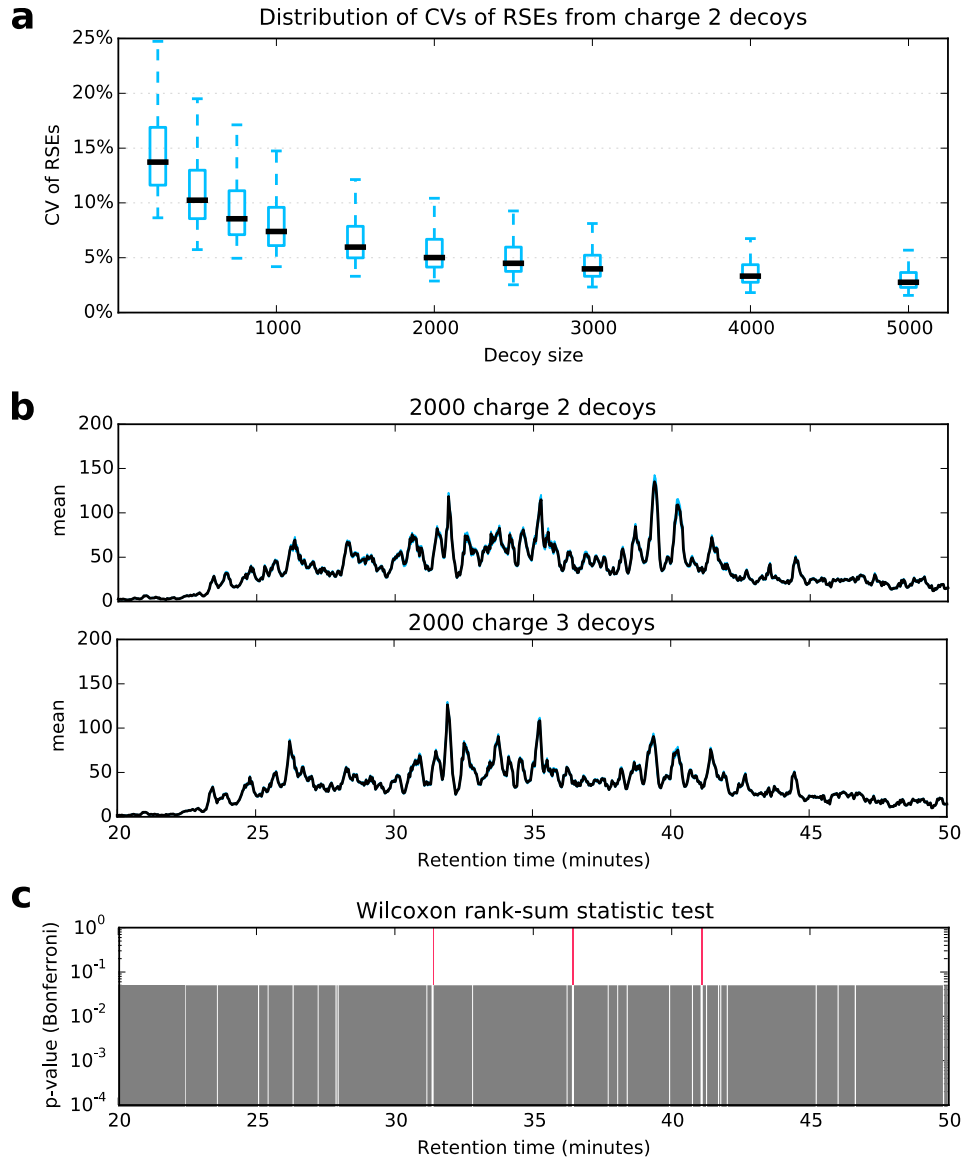


Relative concentration values of 248 plasma proteins are taken from the literature.

(Source: Leigh Anderson, The Plasma Proteome Institute, Washington, DC, USA, modified from ref *Mol. Cell Proteomics* 1, 845–847, 2002.) Color of the dot represents the number of peptides unique to the protein or only shared by its isoforms in the DIA plasma library.

Note that some literature values are measurement for protein complex or specific fragments of the protein (e.g. values for Prothrombin and Fibrinogen alpha chain), of which the intact protein concentration could be higher.

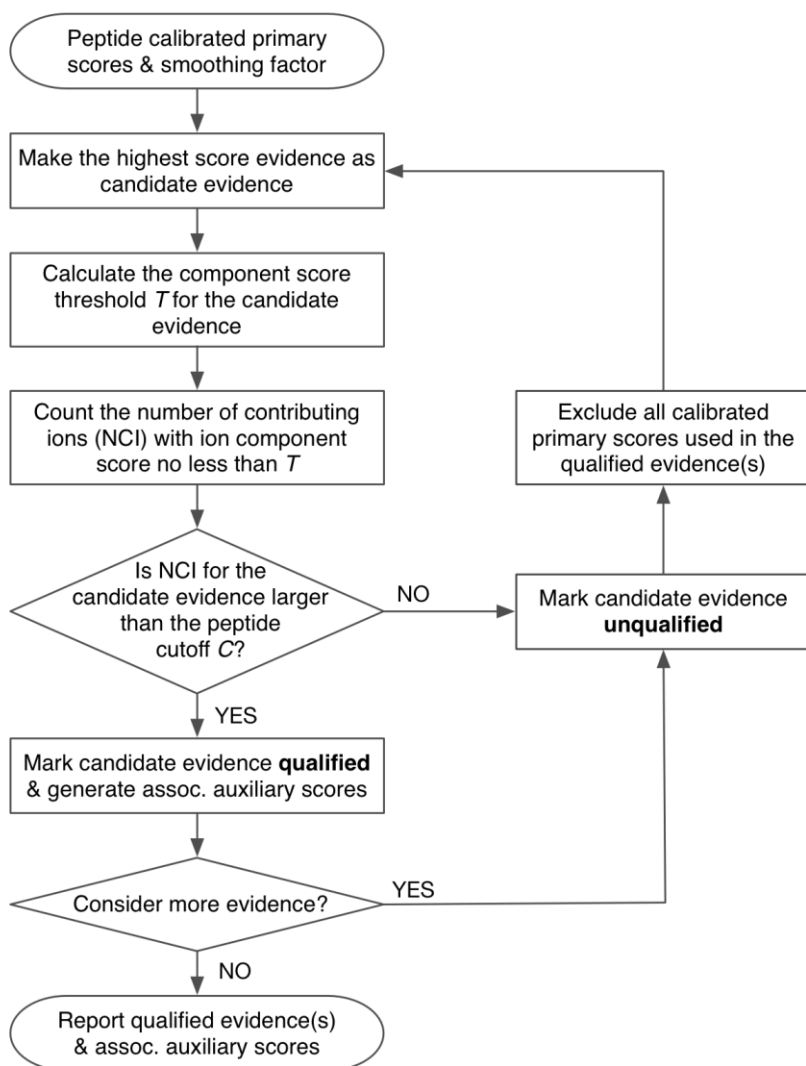
Supplementary Figure 3 – Assessment of background scores estimation with 1,000 random sampling



(a) Boxplot shows the distribution of 2,185 CVs of the RSEs from 1,000 random sampling at each decoy size. (b) The estimated background scores with 2,000 charge 2 and 2,000 charge 3 decoys for 2,185 MS/MS spectra presented over retention time. Black lines trace the median of the decoy means from 1,000 estimations by random sampling and the blue shades are segments between the 25th and 75th percentiles. (c) Bonferroni corrected p-values from Wilcoxon rank-sum tests between the 1,000 estimations using either 2,000

charge 2 or 2,000 charge 3 decoys for individual spectrum. Grey lines indicated the p-value is smaller than 0.05 and therefore rejected the null hypothesis.

Supplementary Figure 4 – Evidence qualifying procedure in PECAN



An evidence of detection (abbr. evidence) for a query peptide p at the time t is the average of the calibrated primary scores from a short period of retention time (see Methods), centered at the time t . Following this flowchart, PECAN reports a user-defined number of qualified evidence(s) that are calculated from primary scores which have never been used to calculate other qualified evidences(s).

Supplementary Tables

Supplementary Table 1 – Variant-specific peptides

Feature identifier	Accession	Variant	dbSNP	Peptide
VAR_025657	P00450	Asp544Glu	rs701753	MYSSAVEPTKDIFTGLIGPMK
VAR_025657	P00450	Asp544Glu	rs701753	MYSSAVEPTK
VAR_006711	P00734	Glu200Lys	rs62623459	SKGSSVNLSPPLEQCVPR
VAR_011781	P00734	Thr165Met	rs5896	NPDSSTMGPWCYTDDPTVR
VAR_005294	P00738	Asn129Asp	rs199926732	TEGDGVYTLNDEK
VAR_006580	P00740	Asn283Asp	-	ITVVAGEHDIEETEHEQK
VAR_006533	P00740	Glu54Gly	-	LEGFVQGNLER
VAR_017356	P00740	Ile344Leu	-	EYTNLFLK
VAR_011779	P00747	Ile46Arg	rs1049573	EECAAKCEEDEEFTCR
VAR_014336	P00748	Ala207Pro	rs17876030	LCHCPVGYTGPFCDDVTK
VAR_016277	P00751	Lys565Glu	rs4151659	EEAGIPEFYDYDVALIK
VAR_027451	P01008	Cys32Arg	-	HGSPVDICTAKPR
VAR_027452	P01008	Tyr95Cys	-	FATTFCQHLADSK
VAR_006995	P01009	Gln180Glu	-	EINDYVEK
VAR_006996	P01009	Glu228Lys	rs199422208	DTKEEDFHVDQVTTVK
VAR_007010	P01009	Glu400Asp	rs1303	FNKPFVFLMIDQNTK
VAR_026820	P01023	Asn639Asp	rs226405	DLTGFPGLNDQDDEDCINR
VAR_063217	P01024	Asp1115Asn	rs121909585	QKPNGVFQEDAPVIHQEMIGGLR
VAR_063219	P01024	Gln1161Lys	-	DICEEKVNSLPGSITK
VAR_048853	P01042	Asp430Glu	rs5030084	RHEWGHEK
VAR_073349	P01602	Lys72Asp	-	LLIYDASSLESGVPSR
VAR_003897	P01834	Val83Leu	-	LYACEVTHQGLSSPVTK
VAR_003897	P01834	Val83Leu	-	HKLYACEVTHQGLSSPVTK
VAR_068700	P01860	Asn245Asp	-	VVSVLTVLHQDWLDGK
VAR_068700	P01860	Asn245Asp	-	VVSVLTVLHQDWLDGKEYK
VAR_003903	P01871	Gly191Ser	-	ESDWLSQSMFTCR
VAR_014602	P01876	Glu176Asp	rs1407	DASGVTFWTPSSGKSAVQGPPDR
VAR_014602	P01876	Glu176Asp	rs1407	SAVQGPPDR
VAR_003102	P02042	Gly25Asp	rs34460332	VNVDAVDGEALGR
VAR_003103	P02042	Gly26Asp	rs34389944	VNVDAVGDEALGR
VAR_000612	P02647	Ala119Asp	-	DKVQPYLDDFQK
VAR_000617	P02647	Glu134Lys	-	WQKEMELR
VAR_000618	P02647	Glu160Lys	rs121912718	LQEKLSPLGEEMR
VAR_000625	P02647	Glu222Lys	rs121912717	ATKHLSTLSEK
VAR_000615	P02647	Lys131Met	rs4882	MWQEEMELR
VAR_000649	P02649	Gln99Lys	-	SELEEKLTPVAEETR
VAR_013093	P02675	Pro265Leu	rs6054	KGGETSEMYLIQPDSSVK

VAR_013093	P02675	Pro265Leu	rs6054	GGETSEMYLIQPDSSVK
VAR_014170	P02679	Gly191Arg	rs6063	LYFIKPLK
VAR_036018	P02751	Asp940Asn	rs752106647	VNVIPVNLPGEHGQR
VAR_061486	P02751	Val2170Ile	rs1250209	GATYNIIVEALK
VAR_061486	P02751	Val2170Ile	rs1250209	GATYNIIVEALKDQQR
VAR_007591	P02766	Arg124Cys	rs745834030	CYTIAALLSPYSYSTTAVVTNPKE
VAR_038967	P02766	Asp58Ala	-	KAAADTWEPFASGK
VAR_038968	P02766	Asp58Val	-	AAVDTWEPFASGK
VAR_007585	P02766	Glu109Gln	rs121918082	ALGISPFHQHAEVVFTANDSGPR
VAR_010659	P02766	Glu109Lys	-	ALGISPFHKHAEVVFTANDSGPR
VAR_038976	P02766	Glu74Lys	-	TSESGKLHGLTTEEEFVEGIYK
VAR_007583	P02766	Ile104Asn	-	ALGNSPFHEHAEVVFTANDSGPR
VAR_038985	P02766	Ile127Met	-	YTMAALLSPYSYSTTAVVTNPKE
VAR_007576	P02766	Ile88Leu	rs121918085	TSESGELHGLTTEEEFVEGLYK
VAR_007594	P02766	Leu131Met	rs121918073	YTIAALMSPYSYSTTAVVTNPKE
VAR_007570	P02766	Leu78His	rs121918069	TSESGELHGHHTTEEEFVEGIYK
VAR_038961	P02766	Ser43Asn	-	VLDVRGNPAINVAVHVFR
VAR_007595	P02766	Tyr134Cys	rs121918075	YTIAALLSPCSYSTTAVVTNPKE
VAR_000527	P02768	Asp389His	rs77187142	CCAAHPHECYAK
VAR_000530	P02768	Asp399Asn	rs77514449	VFNEFKPLVEEPQNLIK
VAR_000542	P02768	Asp587Asn	rs76587671	ADNKETCFAEEGK
VAR_000508	P02768	Asp87Asn	rs78574148	TCVADESAENCNK
VAR_000509	P02768	Glu106Lys	rs80296402	KTYGEMADCCAK
VAR_000509	P02768	Glu106Lys	rs80296402	TYGEMADCCAK
VAR_000511	P02768	Glu143Lys	rs75522063	LVRPKVDVMCTAFHDNEETFLK
VAR_000511	P02768	Glu143Lys	rs75522063	VDVMCTAFHDNEETFLK
VAR_000511	P02768	Glu143Lys	rs75522063	VDVMCTAFHDNEETFLKK
VAR_000526	P02768	Glu382Lys	rs75791663	KCCAAADPHECYAK
VAR_000532	P02768	Glu400Gln	rs79047363	VFDQFKPLVEEPQNLIK
VAR_000531	P02768	Glu400Lys	rs79047363	VFDKFKPLVEEPQNLIK
VAR_000531	P02768	Glu400Lys	rs79047363	FKPLVEEPQNLIK
VAR_000533	P02768	Glu406Lys	rs76483862	EPQNLIK
VAR_014294	P02768	Glu420Lys	-	QNCELKQLGEYK
VAR_000536	P02768	Glu525Lys	rs75523493	KFNAETFTFHADICTLSEK
VAR_000536	P02768	Glu525Lys	rs75523493	FNAETFTFHADICTLSEK
VAR_000537	P02768	Glu529Lys	rs74826639	EFNAKTFTFHADICTLSEK
VAR_000537	P02768	Glu529Lys	rs74826639	TFTFHADICTLSEK
VAR_000543	P02768	Glu589Lys	rs75709682	KTCFAEEGK
VAR_000512	P02768	His152Arg	rs80095457	LVRPEVDVMCTAFRDNEETFLK
VAR_000515	P02768	Lys249Gln	rs79804069	FPQAEFAEVSK
VAR_013016	P02768	Lys383Asn	rs75069738	LAKTYETTLENCCAAADPHECYAK
VAR_013012	P02768	Val146Glu	rs77752336	LVRPEVDEMCTAFHDNEETFLK
VAR_013012	P02768	Val146Glu	rs77752336	LVRPEVDEMCTAFHDNEETFLKK
VAR_058199	P02787	Ile448Val	rs2692696	SDNCEDTPEAGYFAVAVVK

VAR_058199	P02787	Ile448Val	rs2692696	SDNCEDTPEAGYFAVAVVKK
VAR_012000	P02787	Pro589Ser	rs1049296	SVEEYANCHLAR
VAR_012000	P02787	Pro589Ser	rs1049296	DYELLCLDGTRK
VAR_012000	P02787	Pro589Ser	rs1049296	KSVEEYANCHLAR
VAR_016286	P03952	Arg560Gln	rs4253325	ITQQMVCAGYK
VAR_059582	P04114	Ile2313Val	rs584542	INDVLEHVK
VAR_029342	P04114	Pro877Leu	rs12714097	LEVANMQAELVAK
VAR_061558	P04114	Tyr1422Cys	rs568413	NTFTLSCDGLR
VAR_024429	P04196	Asn493Ile	rs1042464	HPLKPDIQFPQSVSESCPGK
VAR_018369	P04217	His52Arg	rs893184	LETPDFQLFK
VAR_038628	P04264	Ala454Ser	rs17678945	LNDLEDALQQSK
VAR_000627	P06727	Glu44Lys	-	KAVEHLQK
VAR_046821	P07225	Cys121Tyr	-	SCVNAIPDQYSPLPCNEDGYMSCK
VAR_033800	P07357	Asp458Asn	rs17114555	YNPVVINFEMQPIHEVLR
VAR_011889	P07357	Gln93Lys	rs652785	KAQCGQDFQCK
VAR_011892	P07357	Glu561Gln	rs1342440	QCDNPAPQNGGASCPGR
VAR_019406	P08603	Cys959Tyr	-	YFEGFGIDGPAIAK
VAR_025093	P08603	Ser890Ile	rs515299	SSQEIIAHGTKLSYTCEGGFR
VAR_023836	P08603	Val62Ile	rs800292	SLGNIIMVCR
VAR_072438	P08779	Asn125Asp	rs58608173	VTMQNLDDR
VAR_069154	P0C0L4	Leu141Val	rs9296005	GHVFLQTDQPIYNPGQR
VAR_069154	P0C0L4	Leu141Val	rs9296005	RGHVFLQTDQPIYNPGQR
VAR_069160	P0C0L5	Pro478Leu	-	LTVAAPPSGGPGFLSIER
VAR_033799	P10643	Thr587Pro	rs13157656	DGFVQDEGPMFPVGK
VAR_001214	P12259	Lys858Arg	rs4524	LLSLGAGEFR
VAR_069914	P12814	Glu225Lys	rs387907350	MLDAKDIVGTARPDEK
VAR_017475	P14136	Glu362Asp	rs28932768	LALDIDIATYR
VAR_050173	P19652	Gly141Arg	rs12685968	NWRLSFYADKPETTK
VAR_004020	P19827	Gln595Arg	rs1042779	MSLDYGFVTPLTSMIR
VAR_044226	P35579	Lys910Gln	rs554332083	QQELEEEICHDLEAR
VAR_007639	P49747	Asp518Asn	-	INVCPEAEVTLTDFR
VAR_012857	P68032	Glu101Lys	rs193922680	VAPKEHPTLLTEAPLNPK
VAR_062436	P68133	Ile77Leu	-	YPIEHGLITNWDDMEK
VAR_062427	P68133	Pro40Leu	-	AVFPSIVGR
VAR_003031	P68871	Asn109Lys	rs34933751	VLVCVLAHHFGK
VAR_003077	P68871	Asn140Asp	rs33910475	VVAGVADALAHK
VAR_002886	P68871	Asn20Asp	rs34866629	VDVDEVGGEALGR
VAR_002887	P68871	Asn20Lys	rs63750840	VDEVGGEALGR
VAR_002891	P68871	Asp22Asn	rs33950093	VNVNEVGGEALGR
VAR_002890	P68871	Asp22Gly	rs33977536	VNVGEVGGEALGR
VAR_003058	P68871	Gln128Glu	rs33971634	EFTPPVEAAYQK
VAR_002927	P68871	Gln40Glu	rs76728603	LLVVPWTER
VAR_003048	P68871	Glu122Gln	rs33946267	QFTPPVQAAYQK
VAR_003049	P68871	Glu122Lys	rs33946267	KFTPPVQAAYQK

VAR_002897	P68871	Glu23Gln	rs33959855	VNVDQVGGEALGR
VAR_002793	P69905	Asp75Asn	rs281864857	VADALTNAVAHVNDMPNALSALS DL HAHK
VAR_034541	Q13748	Val75Leu	rs36215077	AVFVDLEPTVLDEV R
VAR_027870	Q14624	Gln669Leu	rs2276814	LLGLPGPPDVPDHAAYHPFR
VAR_014761	Q16610	Gly415Ser	rs13294	DILTIDIS R
VAR_032337	Q6UXB8	Thr50Pro	rs1405069	AQVSPPASDMLHMR
VAR_049062	Q9UGM5	Lys360Arg	rs7999	LVVLPFPR

Supplementary Table 2 – Auxiliary scores for qualified evidence of detection

Feature	Level	Description
peak score	fragment	Average of pre-calibrated primary scores from a short period of time centered at the retention time t for the evidence
peak calibrated score	fragment	Average of calibrated primary scores from a short period of time centered at the retention time t for the evidence (i.e. $E_p(t)$, the evidence of detection for peptide p at time t)
peak weighted score	fragment	Average weighted score of pre-calibrated primary scores from a short period of time centered at t , where each fragment ion contribution is weighted by multiplied with its m/z value
peak Z score	fragment	Average of standardized calibrated primary scores from a short period of time centered at the retention time t for the evidence, where each calibrated primary score is standardized with the mean and standard deviation of the 2,000 decoy scores of the same precursor charge state
spectra norm	fragment	Average of magnitudes of MS2 spectrum within a short period of time centered at the evidence retention time, where each magnitude is calculated as the Euclidean length of spectrum with square root of the intensities.
NCI	fragment	Number of contributing ions (CIs)
rank	fragment	Rank of the evidence relative to other qualified evidences (if any) for the query peptide
delta Sn	fragment	Normalized delta "peak calibrated score" of the evidence to the next qualified evidence
CI mass error mean	fragment	Mean of the weighted mass errors in ppm from the contributing ions (CI), where the mass error of each CI is weighted by the observed intensity
CI mass error variance	fragment	Variance of the weighted mass errors in ppm from the contributing ions (CI), where the mass error of each CI is weighted by the observed intensity
similarity	fragment	Average cosine similarity of the observed spectra to the peptide scoring vector, where the observed spectra are MS/MS spectra from a short period of time centered at the evidence time t
sampled times	fragment	Number of MS/MS spectra from a short period of time centered at the retention time t of the evidence
retention time	fragment	Midpoint retention time t of the evidence
Average idotp	precursor	Average isotopic dot product score between expected and observed isotopic envelope distributions from MS1 spectra of a short period of time centered at the evidence time t
Midpoint idotp	precursor	Isotopic dot product score between expected and observed isotopic envelope distributions from MS1 spectrum at the center time t of the evidence
precursor mass error mean	precursor	Mean of the weighted mass errors in ppm from the precursor ions, where the mass error of each precursor ion is weighted by the observed intensity
precursor mass error variance	precursor	Variance of the weighted mass errors in ppm from the precursor ions, where the mass error of each precursor ion is weighted by the observed intensity
peptide length	peptide	Numbers of amino acid from the query peptide

precursor charge state	peptide	Charge state of the query peptide precursor
---------------------------	---------	---------------------------------------------

Supplementary Table 3 – Direct links for downloading the raw files

Dataset Name	Chorus ID	Link
SRM validation of IVTT proteins	2427	https://chorusproject.org/anonymous/download/experiment/4846597907291871276
HeLa datasets part I: DDA	2448	https://chorusproject.org/anonymous/download/experiment/-2822210361803919543
HeLa datasets part II: DIA	2449	https://chorusproject.org/anonymous/download/experiment/1929128726775705417
DIA plasma library	2655	https://chorusproject.org/anonymous/download/experiment/-3803766532162238398

Supplementary Notes

Supplementary Note 1 – Assessment of background scores estimation

Background scores estimation is a key component to PECAN scoring. As discussed in the main manuscript, MS/MS spectra acquired with DIA contain many peptide-like fragment ions. Thus, any peptide could score none-zero against the same MS/MS spectrum. To estimate how high on average a peptide score can be achieved merely by chance with a dataset, PECAN calculates estimated background scores represented by the arithmetic means of thousands of decoy peptides over time. These decoys are generated from shuffling a random selection of proteolytic peptides from the background proteome databases, typically the protein sequence database of the targeted species when analyzing complex sample (see Supplementary Not 6 – FAQ).

The approach PECAN uses to estimate a background score for individual spectrum is analogous to estimating the population mean using a random sample. Because even with a strict proteolytic rule (e.g. fully trypsin digestion), calculating the population mean from all possible proteolytic peptides and precursor ions for every MS/MS spectrum is computationally expensive. For example, there are $\sim 10^{10}$ possible unmodified peptides with C-terminal arginine or lysine that could generate charge 2 precursor ions between 500-505 m/z . In light of this, we adopted the standard practice of estimating the population mean using a random sample.

To determine the sample size N (i.e. number of decoys) for background scores estimation, we selected ten different sizes and evaluate the resulting estimate with relative standard error of the mean (RSE), a standard metric indicates how far the estimate is likely to be from the true population mean expressed as a fraction of the estimate. In addition, to account for the sampling effect, for each sample size we performed 1,000 estimations, resulting 1,000 RSEs for every spectrum (Supplementary Fig. 3a). In this experiment, we used data from one isolation window (500-505 m/z) of a mouse DIA dataset that contains 2,185 MS/MS spectra between retention time 20-50 minutes, where most of the peptides were eluted. Charge 2 decoys were generated from random sampling the corresponding

size of possible tryptic peptides without replacement from the mouse Swiss-Prot database. In one estimation, a set of N decoys were generated to calculate 2,185 sample means for 2,185 spectra, followed by 2,185 RSEs. According to the central limit theorem, both the sample means and the RSEs from 1,000 random sampling should be normally distributed. In light of this, to demonstrate sampling effect and evaluate the robustness of the estimation, we calculated the coefficient of variation (CV) of the 1,000 RSEs for individual spectra. Overall, the CVs of the 1,000 RSEs across the data decreased as the sample size increased (Supplementary Fig. 3a). At sample size 2,000, the RSEs of more than 75% of the 2,185 spectra varied less than 7% CV. Thus, we chose decoy size 2,000 for background score estimation throughout the current study.

Next, we wanted to determine if background scores should be charge state dependent. We used Wilcoxon rank-sum test with the null hypothesis that the underlying score distribution for each MS/MS spectrum from charge 2 and charge 3 decoys are identical (Supplementary Fig. 3b). At decoy size of 2,000, only 30 of 2,185 spectra tested with Bonferroni corrected p-value ≥ 0.05 and therefore failed to reject the null hypothesis. This number was further reduced when we increased the decoy size (data not shown). This results demonstrated that the majority of the underlying background score distribution from charge 2 and charge 3 decoys are not identical, and in cases where the two distributions appeared to be identical it was likely an effect of sample size. Thus, PECAN estimates background scores in a charge state dependent fashion.

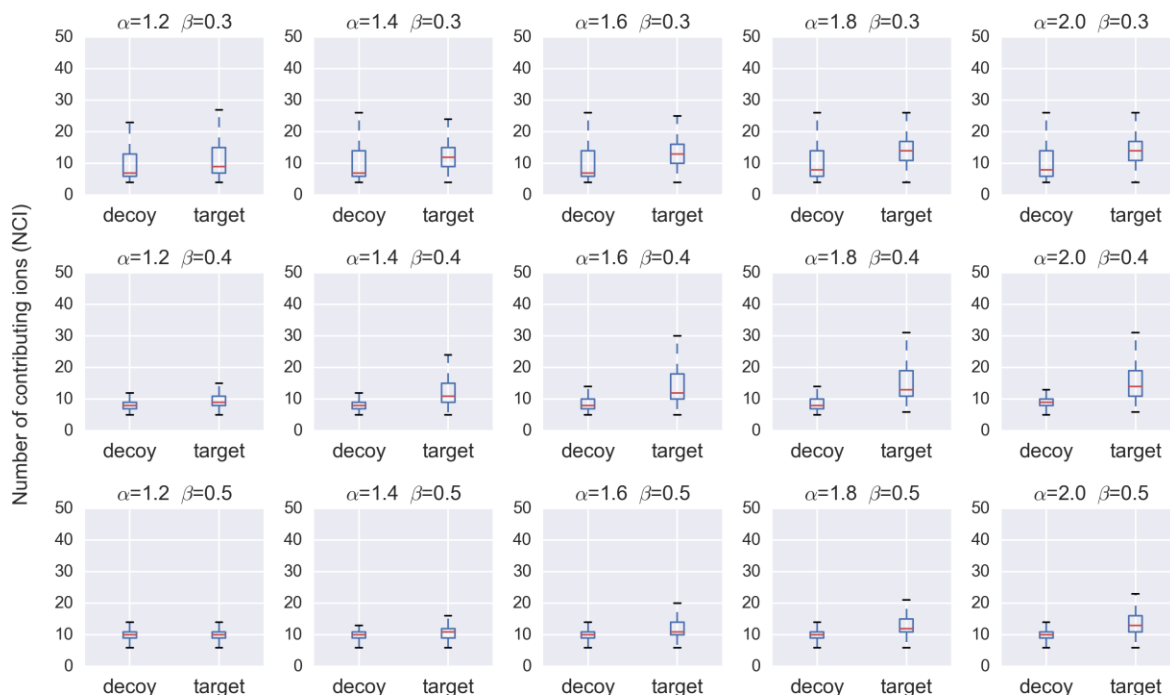
Supplementary Note 2 – Hyperparameters determination for the evidence qualifying procedure

PECAN uses empirical criteria during evidence qualifying procedure to disqualify evidences of detection whose scores are predominantly contributed by a small number of fragment ions, suggesting that the score could be resulting from interference of a few high abundance ions rather than a collaboration of multiple fragment ions. Two hyperparameters α and β are used to set the criteria as described in the main manuscript.

To determine the hyperparameter α and β , we used a *S. cerevisiae* lysate DIA dataset, acquired on a Q-Exactive using a 10- m/z -wide isolation window DIA approach in which the mass range from 500 to 700 m/z is analyzed with twenty non-overlapping 10- m/z wide isolation window targeted MS/MS scans. This dataset contained 6 biological replicates; each included manually curated boundaries of chromatographic peaks from 204 peptides verified by DDA identification. A total of 1,224 peak boundaries were used as reference for the following test (available at Panorama Public).

We first looked at the NCI distribution of PECAN reported evidences resulted from various combinations of α and β (shown below). Overall, as the α increased, the median of NCI distribution also increased. This is expected because the incensement of α decreased the component score threshold of each evidence. As a result, with a lower component score threshold, more fragment ions were considered “contributing ions” for passing the threshold. On the other hand, as the β increased, the range of NCI distribution became tighter, especially for decoys. Because β controls the threshold of NCI required for an evidence to be qualified, larger β favors evidences with more uniformly distributed component contributions. However, the larger the β is the less sensitive the evidence qualifying procedure is. Finding the balance between α and β is key to the sensitivity and specificity of the procedure.

NCI distributions with various hyperparameter combinations

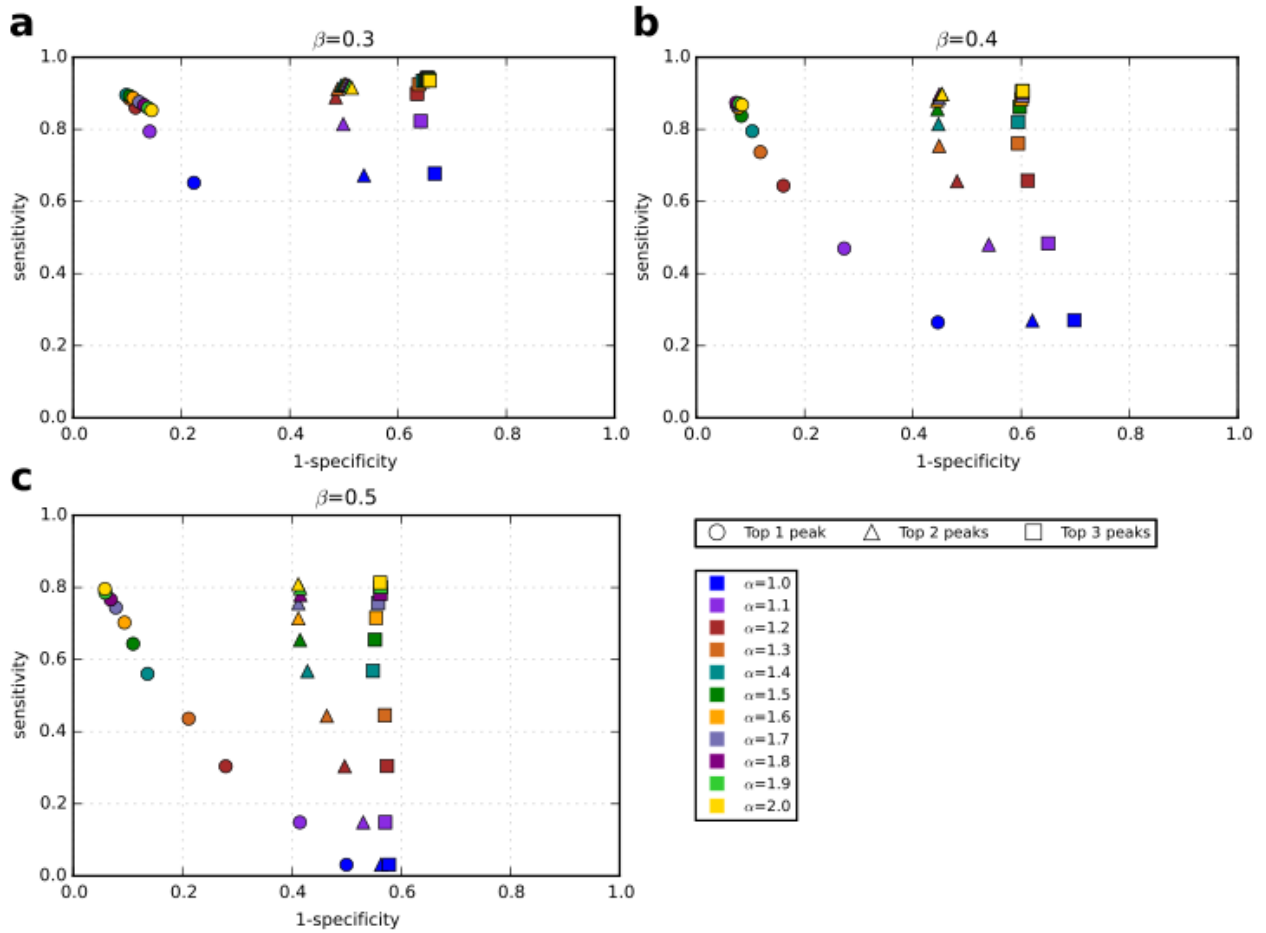


Box plots show the NCI distribution of PECAN reported top 1 evidences of detection from 12 representative sets of α and β . Both target and decoy evidences reported by PECAN are included without any FDR control.

With different α and β , we evaluated the performance of evidence qualifying procedure by comparing the reference peak boundaries to the retention time of PECAN reported evidences when considering top 1, top 2, or top 3 evidence(s) for each query peptide. A reported evidence was classified as correct if the reported retention time (i.e. center time of the evidence) had fallen between the reference peak boundaries of the query peptide. We defined sensitivity to be the number of peptides with one correct evidences over the total number of query peptides, and specificity to be the number of correct evidences over the total number of reported evidences. As expected by these definitions, we observed that at a given set of α and β , specificity dropped significantly when PECAN reported top 2 or top 3 evidences per peptide with minimum sensitivity gains compared to reported only the

top 1 evidence (shown below). This result indicates that the calibrated primary score PECAN used to rank the candidate evidences of detection for each peptide was effective so that rarely the second or third best evidence were correct. Together, with $\beta=0.4$ and $\alpha=1.8$ PECAN resulted the best balance between sensitivity and specificity determined by area under the curve when consider only the top 1 evidence (shown below panel b). This set of α and β values were used throughout the current study.

Performance of the evidence qualifying procedure with different hyperparameters



Sensitivity and specificity of the qualifying procedure when reported top 1, top 2, or top3 qualified evidence(s) of detection with $1.0 \leq \alpha \leq 2.0$ α and $\beta = 0.3$ (a), $\beta = 0.4$ (b), or $\beta = 0.4$ (c). At any given set of α and β , the sensitivity gains were minimum when reporting top 2 or top 3 qualified evidences compared to only reporting the top 1 qualified evidence,

indicating that the primary score used to rank the qualified evidences of detection for query peptides were effective.

Supplementary Note 3 – Decoy generation

PECAN uses two types of decoys: one for background scores estimation and the other for target-decoy paradigm. Decoy peptides in PECAN are generated by Fisher-Yates shuffling a reference proteolytic peptide while keeping the proteolytic site (e.g. C-terminal R and K for trypsin). In all cases, a decoy is invalid if it is present in either the list of query (target) peptides or the background proteome.

For background scores estimation, the background proteome is used to seed for decoy generation (Supplementary Note 1). A new decoy will be generated with the same reference peptide until either a valid decoy has been generated or three attempts has been made. In case of no valid decoy after three attempts, PECAN will shuffle the reference sequence without maintaining the proteolytic site. This tri-shuffling strategy is to ensure the expected number of valid decoys is successfully generated so that the relative standard error of mean (RSEs) is not underestimated (Supplementary Fig. 3a).

For use of target-decoy paradigm, the list of query (target) peptides is used to seed for decoy generation. In PECAN, because the size of the target list could vary from a few thousands to several millions, it is essential to ensure that the resulting decoys properly represent the “null” population. As shown in the literature on spectral library searching¹², decoys generated from a smaller set of targets could be biased toward the reference targets, and thus fail to properly represent the null. In this case, the FDR could be overestimated because of the bias in the target list. For example, if only peptides known to be abundantly present in the sample is queried, the high similarity decoys generated from these targets would likely be biased towards the target distribution and poorly portray the true null. Typically, a large query, such as querying all gene products of a species, is less likely to have such bias because the query list consists of a mixture of present and absent peptides. In contrast, a smaller query, such as querying only the peptides from a metabolic pathway, could have such target bias effect. To account for this potential effect associated

¹ Lam, H., Deutsch, E. W., & Aebersold, R. Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics. *J. Proteome Res.* 9, 605–610 (2010).

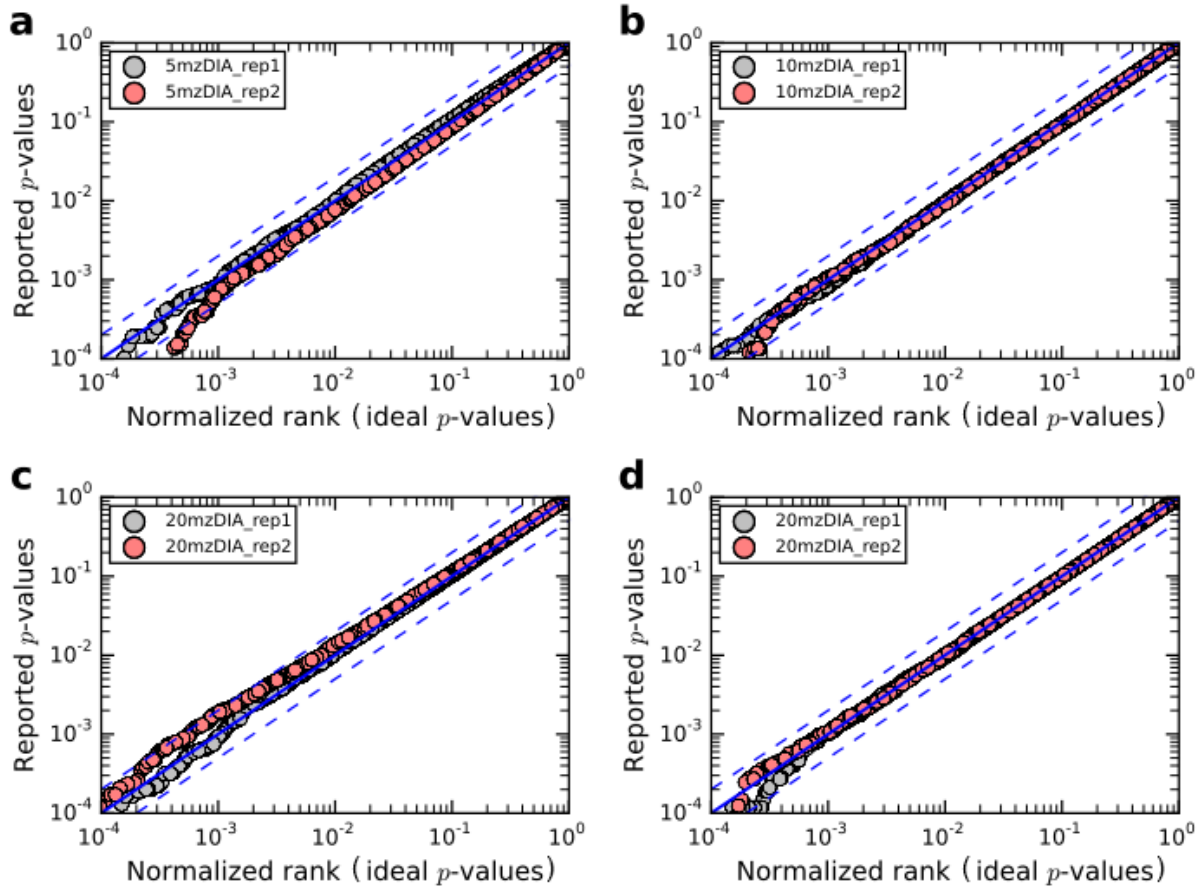
² Ahrné, E. *et al.* An improved method for the construction of decoy peptide MS/MS spectra suitable for the accurate estimation of false discovery rates. *Proteomics.* 20, 4085-4095 (2011)

with variable sizes of query, a decoy here is further validated by the fragment similarity to its reference. Upon generation, a decoy is also invalid if it shares more than 40 % of the theoretical fragment ion m/z values with its reference. A new decoy will be generated by shuffling the same reference until either a valid decoy has been generated from the reference or ten attempts has been made. In case of no valid decoy after ten attempts, the decoy with the least shared theoretical fragment ion m/z values will be used. With this strategy, a decoy will always contain the same amino acid composition, length, molecular weight, and proteolytic site as its reference. Additionally, the similarity-check further ensures that a decoy fragmentation pattern is diverse from its target if possible, thus counterbalance the potential target bias without drawing from the genome sequences.

To evaluate PECAN's decoy strategy (shuffle plus similarity-check), we queried the *E. coli* proteome against HeLa DIA datasets with various DIA isolation schemes. Because neither target nor decoy peptides were present in the sample, the reported target and decoy evidence of detection should not be distinguishable. The results show that PECAN's decoy model generates decoys indistinguishable from the query targets, which in this case are true null to the HeLa digest (shown below).

We further compared the PECAN decoy model with the tri-shuffling model and the reverse sequence model by querying the human UniProt proteome against the plasma library DIA dataset. For all three decoy models, the proteolytic site of each peptide is maintained. As expected, the reverse model yielded the least number of evidence of detection for decoys, largely because that this model only has one chance of generating a valid decoy per target, and thus resulting in a lower number of valid decoys. Because the UniProt human proteome is an unbiased query for the plasma sample, we expect the decoy score distributions from PECAN model to be indistinguishable from the tri-shuffle model. Indeed, two-sample Kolmogorov–Smirnov test results indicate that the primary score distribution of decoys from PECAN model and the tri-shuffling model are from the same distribution, whereas the decoys from the reverse model are not. In summary, when the query is unbiased, the similarity-check in the PECAN model does not result in distinguishable decoys from the tri-shuffling model.

Q-Q plots of reported and ideal p-values with various DIA datasets



Reported p-values are plotted relative to an ideal, uniform distribution of p-values. All p-values were estimated using the Percolator score. The $y = x$ diagonal is indicated by a blue line, and both $y = 2x$ and $y = x/2$ are shown in blue dashed lines. Three HeLa DIA datasets, each containing two technical replicates, were tested: 4xGFP 5mz DIA (a), 2xGFP 10mz DIA (b), and 1xGFP 20mz DIA (c, d). During PECAN analysis, the background proteome used was either the *E. coli* Swiss-Prot protein sequence database (a, b, c), or the human Swiss-Prot protein sequence database (d).

	shuffle + similarity	tri-shuffle	reverse
Number of decoys reported with evidence	339,709	339,431	331,094
p-value of K-S test with tri-shuffle	0.6977	-	2.41e-05
p-value of K-S test with reverse	1.66E-04	2.41e-05	-

Supplementary Note 4 – Select proteins and peptides for IVTT SRM

Ninety-one peptides were selected for the 16 GST-fusion proteins based on a preliminary analysis of PECAN during its early development. The proteins and peptides were selected based on the preliminary PECAN results from the 4xGPF HeLa DIA data acquired with 5m/z-wide isolation windows, and on the Comet results from 4xGPF HeLa DDA data.

First, tryptic peptides with up to one missed cleavage from the 8,207 GST-fusion- protein database were queried against DIA data by PECAN. DDA data was analyzed by Comet using the same database and up to one missed cleavage was allowed. The detected peptides, both reported at Percolator q-value<0.01, were compared and mapped to the proteins in the GST-fusion-protein database. From a random order, the first 16 proteins³ with at least more than 3 additional peptides detected by PECAN-DIA compared to Comet-DDA, and with at most 1 peptide identified by Comet-DDA were selected for IVTT synthesis.

As mentioned in the main manuscript, the SRM assay development described above was done with peptides detected by an earlier version of PECAN. Since then, minor adjustments were made and additional features, such as hyperparameters alpha and beta, were added to PECAN. The earlier version of PECAN was only used in selecting the peptides for IVTT and SRM. All the validation and comparison of PECAN detection in this manuscript and supplementary were performed with PECAN (v 0.9.9).

³ During the culturing step, one of the 16 clones (library well ID: HsxXG003443-A06) did not grow to the desired O.D. We replaced that protein with one that passed all of the aforementioned criteria, but had already been synthesized in the lab (HsxXG006208-E04).

Supplementary Note 5 – Deep gas-phase fractionation DDA

As a reference for deep gas-phase fractionation (GPF) DIA analysis, we also analyzed the DDA data acquired with matching GPF settings. We searched the 1xGPF, 2xGPF and 4xGPF DDA data with Comet and used Percolator and Fido to report peptide and protein identification at q -value < 0.01, respectively. With different GPF settings, DDA should sample in various depths using the same top-20 method because each fractionation focused on a various width of precursor m/z range.⁴ From the 1xGPF, 2xGPF and 4xGPF DDA data, we identified 5,934, 5,915, and 6,221 unique peptides, and 1,504, 1,678, and 1,759 protein groups, respectively (shown below panel a).

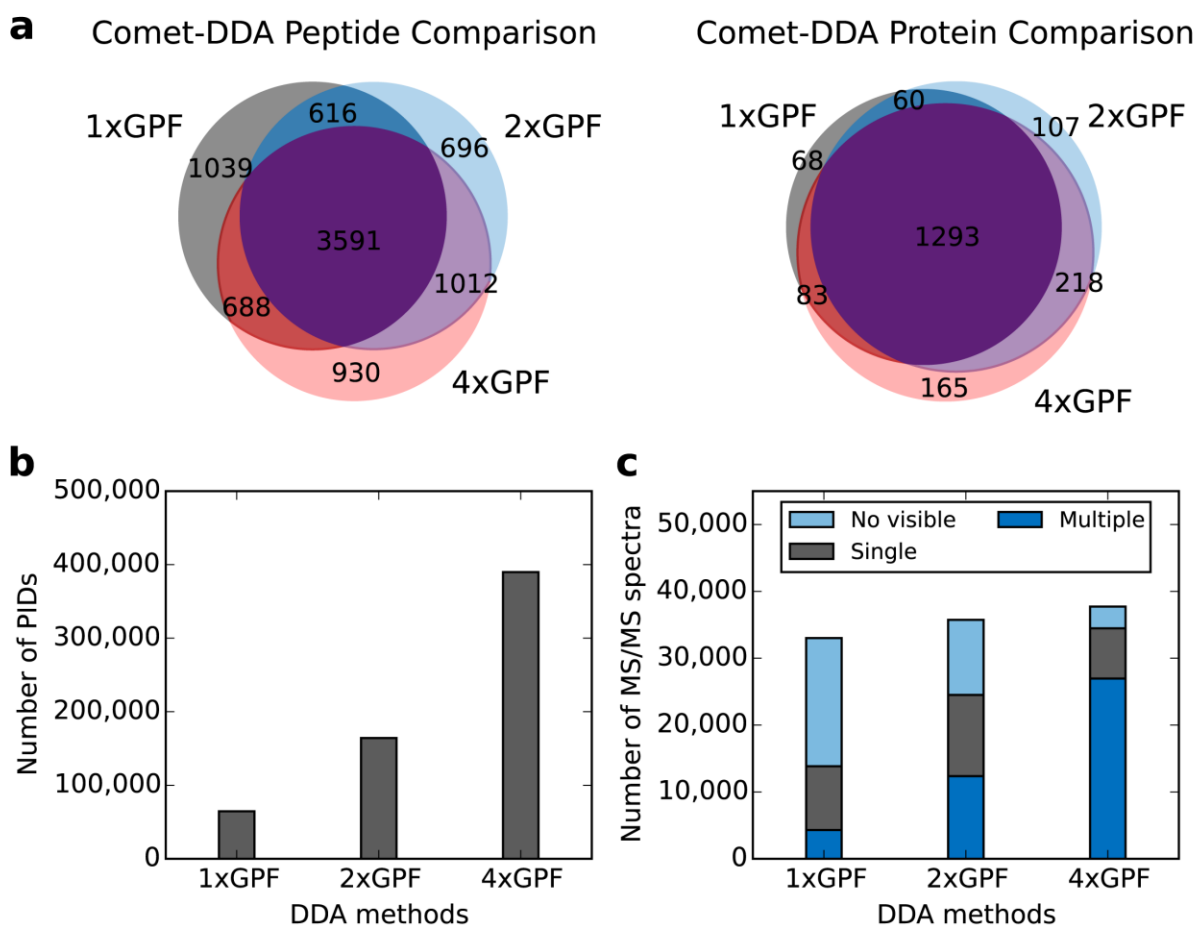
Surprisingly, when we compared 4xGPF to 1xGPF DDA data, only 14 % more MS/MS spectra were acquired (shown below panel c). This was unexpected because 4xGPF cost four times the sample and instrument time of what 1xGPF cost so that in each fractionation DDA only needed to sample from one quarter of the precursor range. In addition, with an Orbitrap mass analyzer, reducing the ion variety for MS1 analysis (i.e. improved MS1 selectivity) by deep GPF should improve the MS1 sensitivity. To test if MS1 sensitivity was improved in deep GPF data, we used Hardklör⁵ (v.2.16) to identify peptide isotopic distributions (PIDs) in the MS1 spectra. As expected, five times more PIDs were identified in 4xGPF than in 1xGPF, indicating that the MS1 sensitivity was greatly improved with deep GPF (shown below panel b). Next, we used Bullseye⁶ (v.1.26) to assign these PIDs within ± 3 seconds in retention time to each MS/MS spectrum. As MS1 signal got more selective from 1xGPF to 4xGPF, significantly higher percentage of MS/MS spectra were assigned with multiple PIDs (shown below panel c). These results indicate that while the DDA method used here was not optimized for the corresponding GPF settings, the sensitivity of MS1 signal was successfully improved by deep GPF.

⁴ Yi, E. C. *et al.* Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* 23, 3205–3216 (2002).

⁵ Hoopmann, M. R., MacCoss, M. J. & Moritz, R. L. Identification of peptide features in precursor spectra using Hardklör and Krönik. *Curr. Protoc. Bioinforma.* 0 13, Unit13.18 (2012).

⁶ Hsieh, E. J., Hoopmann, M. R., MacLean, B. & MacCoss, M. J. Comparison of Database Search Strategies for High Precursor Mass Accuracy MS/MS Data. *J. Proteome Res.* 9, 1138–1143 (2010).

Deep gas phase fractionation revealed more precursor isotope distributions but failed to improve DDA identification due to unoptimized acquisition parameters



(a) Comparison of peptides and proteins identified by Comet from 1xGPF, 2xGPF, and 4xGPF DDA data. (b) Number of peptide isotope distributions (PIDs) identified. (c) Number of MS/MS spectra assigned with no visible, single, or multiple PIDs.

Supplementary Note 6 – Frequently asked questions

Q1. Why did the authors choose to use 8,207 GST-fusion-protein database for validation instead of the more comprehensive database? Why were the number of identifications from HeLa much lower than other studies?

In the “Results – PECAN detection validation”, both the DIA and DDA data sets were analyzed with the GST-fusion-protein database that contains 8,207 proteins. Naturally, the number of peptide identification is much lower compared to other studies that searched HeLa DDA data against other more comprehensive databases, such as the human UniProt Swiss-Prot database (approx. 20,000 reviewed proteins and 42,000 protein isoforms). We chose to use the GST-fusion-protein database for three reasons:

- 1) Using SRM to measure specific peptides from IVTT synthetic proteins is a straightforward and low-cost way to validate peptide detection. Thanks to the DNASU plasmid repository, we have access to full-length cDNA clones for the 8,207 GST-fusion proteins. Because the purpose of this experiment is to validate PECAN detection, we only cared about the peptides from proteins we had access to synthesize full-length proteins ourselves.
- 2) In the recent Nature Methods commentary: “Mass spectrometrists should search only for peptides they care about”⁷, the author demonstrated that removing irrelevant peptides from the database prior to the searching improves statistical power compared to assigning these peptides to spectra and then discarding the matches. Thus, searching with the database of interest (i.e. 8,207 GST-fusion-protein database), we gained statistical power compared to searching the entire UniProt Swiss-Prot database and then filter for the peptides within the database of interest.

⁷ Noble, W. S. Mass spectrometrists should search only for peptides they care about. Nat. Methods 12, 605–608 (2015).

- 3) To evaluate the correctness of PECAN detection, we accepted the detection that agreed with DDA identification and validated a subset of PECAN specific detection with SRM assays.

Therefore, even though the 8,207 GST-fusion-protein database is not the most comprehensive database for a HeLa proteome digest, it contains all the sequences we were interested in for the purpose of validating PECAN detection.

Q2. What is a background proteome? How should I choose a proper background proteome for PECAN? Do I need to consider all possible modifications?

A background proteome is a database provided by the user that contains all expected peptide sequences from the sample. In PECAN, the list of query (target) peptides should only contain peptides of interest. Thus, a background proteome could be different from the list of targets because it may contain sequences from proteins (e.g. keratin) that the user is not interested in. PECAN uses the background proteome for three purposes. First, the background proteome is used to calculate the frequencies of fragment ion m/z values. These frequencies are then used to calculate the weights of each fragment ion relative to a query peptide, so that fragment ions with high frequency m/z values, such as 147.113 (y1-Lysine) and 175.119 (y1-Arginine) for trypsin digestion, are weighted less than those with low frequency m/z values (Online Methods). Second, the background proteome is used to seed for generating decoys that are used in background scores estimation (Supplementary Note 1). Last, PECAN uses the background proteome to make sure that any decoy it generates, in addition to not being in the target list, do not happen to be in the list of expected peptides from the background proteome (Supplementary Note 3).

An ideal background proteome should contain only the peptides expected to be present in the sample. However, as of today, it is still impossible to know the true composition of a proteome, considering the possible post-translational modifications. Fortunately, the fragment ion frequency derived from the background proteome is simply an estimation of how specific one fragment ion is to the peptide relative to other fragment ions. This estimation aims to down weights the high frequency ions, and overlooks the peptide

redundancy in the database and the expected abundance in the sample. Thus, rare events such as nonsynonymous polymorphisms and native modifications do not have high impacts to the estimation.

The guideline to choose the proper background proteome is to consider the majority of the sample without the rare events. For samples from whole cell lysate, tissue, or cell line digest, we recommend the protein sequence database of the corresponding species without native modifications, as native modifications are relatively rare. For samples from an enrichment process such as IMAC, we recommend the protein sequences database of the corresponding species with only the specific modifications, as the unmodified peptides are much less likely to be enriched. Samples from a global labeling process such as SILAC could use the mixed database with both heavy and light.

Q3. What is the importance of each auxiliary score in the percolator SVM? Is there a measure of statistical importance of each score for the different GFP datasets?

There is no measurement of statistical importance for individual auxiliary scores in an SVM. This is because, unlike a method such as logistic regression, which assumes that the underlying data is normally distributed, the SVM is a non-parametric method that makes no assumption about the form or the distribution that generates the data. Without such assumptions, a null model for confidence estimation cannot be analytically derived.

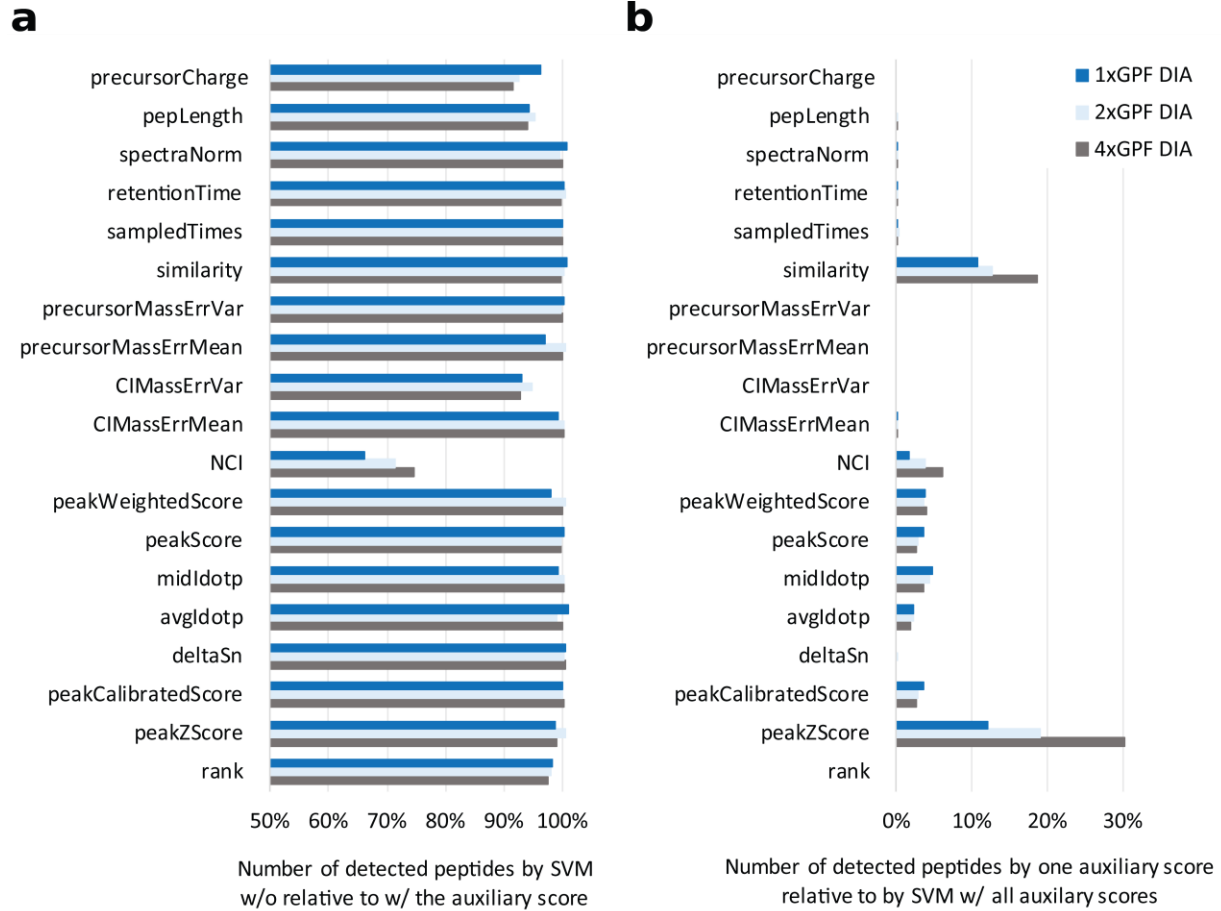
In light of this, practitioners frequently resort to empirical methods to estimate the relative importance of SVM features, in this case the auxiliary scores. This can be done, for example, by deleting one feature from the input and measuring the extent to which this removal affects the performance of the trained classifier. Such methods are admittedly imperfect, both because they do not discriminate between uninformative versus redundant features and because they are conditional on the particular data used for the evaluation. But this type of approach can still provide valuable information.

Unfortunately, this empirical approach still requires a gold standard set of labels against which to evaluate performance. In the case of proteomics, such a gold standard is not

easily obtained. We therefore adopted the standard practice of using an empirical null model based on decoy peptides. With this assumption, we can estimate the importance of an auxiliary score by leaving it out of the SVM. For each auxiliary score, we counted the number of peptides detected by SVM without the score relative to the number of peptides detected with the score in three GPF datasets (shown below panel a). In this leave-one-out analysis, the absence of score NCI had the largest impact on the overall discriminatory power of the SVM.

In addition, we investigated how discriminative an auxiliary score is on its own, independent of the SVM. With the empirical null model based on decoy peptides, we counted the number of peptides detected at q -value < 0.01 by each auxiliary score relative to the number of peptides detected by SVM with all auxiliary scores (shown below panel b). In this leave-one-in analysis, the auxiliary score peakZscore had the highest discriminatory power by itself, averaged out to around 20% of the number of peptides detected by SVM.

Discriminatory power analysis of auxiliary scores



(a) Leave-one-out analysis shows the number of peptides detected with q -value < 0.01 by SVM without the corresponding auxiliary score relative to with the corresponding auxiliary score. (b) Leave-one-in analysis demonstrates the discriminatory power of each auxiliary score on its own with q -value < 0.01 , relative to the power of SVM with all auxiliary scores.

Q4. How does PECAN handle modifications? Is it possible to detect multiple modification forms of one peptide?

PECAN can be used to query modified forms of the peptides. For fixed modifications, such as carbamidomethyl cysteine, the delta mass of the modification is applied globally to the modified residue, including target peptides, peptides in the background proteome, and every decoy generated. For querying peptides with variable modifications, PECAN treats each peptide query independently. PECAN leverages precursor information in the form of auxiliary scores. In the case where multiple modified forms of one peptide have different intact masses, the evidence reported by PECAN for each modified form will have different auxiliary scores, including precursor isotopic dot products, and means and variances of precursor mass error, even if the same group of spectra provides best evidence for more than one modified forms of the peptide. In case of positional isomers, PECAN treats each peptide query independently. Thus, it is possible that multiple isomers could be scored equally high with the same group of spectra.

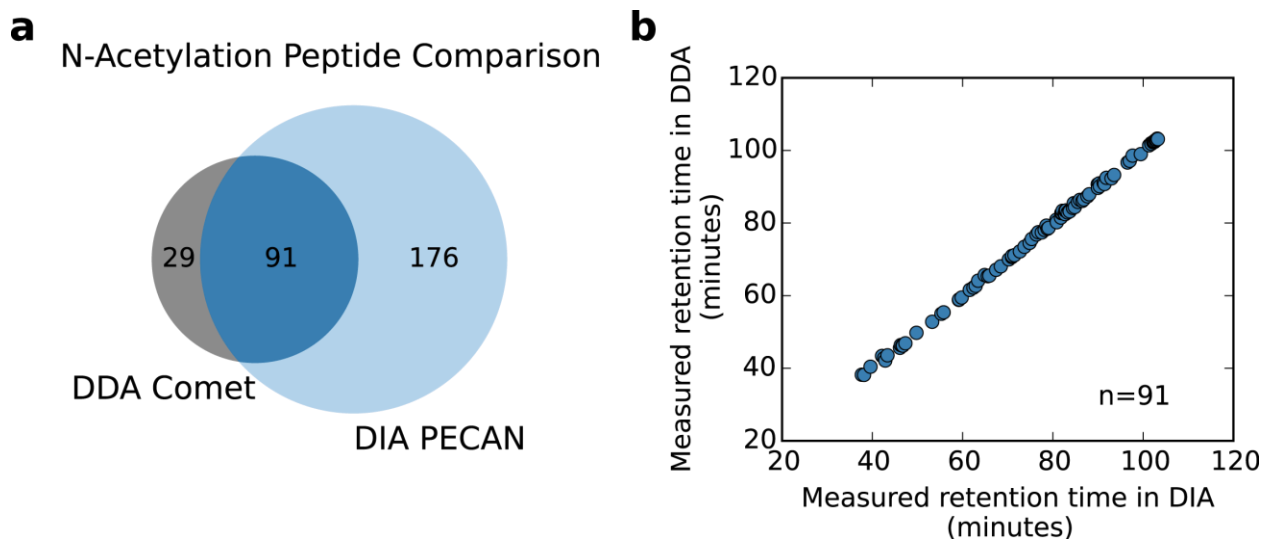
To demonstrate how PECAN performs when considering modifications, we queried the modified peptides from protein *N*-terminal acetylation (i.e. *N*-acetylation) in addition to the unmodified peptides of the human UniProt Swiss-Prot database against the 4xGPF DIA data. PECAN detected 34,958 unique peptides, including 267 peptides from protein *N*-acetylation. In addition, we used Comet to search the 4xGPF DDA data allowing for variable modification of protein *N*-terminal acetylation. Comet identified 15,656 unique peptides including 120 peptides from protein *N*-acetylation (shown below panel a). 91 modified peptides were detected in both methods. The measured retention time of these 91 peptides from DDA and DIA data aligned nicely and thus further confirmed the detection with modification made by PECAN (shown below panel b).

Differentiating modifications from DIA data is a lot more challenging than from DDA data. Depending on the DIA isolation scheme, multiple modification forms of the same peptide could all reside in the same MS2 isolation window. For example, oxidation on a 2+ peptide only has a precursor shift of 8 m/z. The oxidation form and the non-modified form of one peptide could share most of the fragment ions and be measured in the same MS2 scans in

DIA with isolation windows larger than 8 m/z-wide. In this case, one could only distinguish the detection if the MS1 provides strong support preferring one precursor, or if the distinguishing product ions were observed. For this reason, PECAN leverages precursor information when it is available to improve search results and distinguish between the modifications.

Currently, PECAN does not further filter detections if the same group of spectra provided evidence to multiple forms of one peptide. By design, PECAN treats the detection of every peptide independently from others. It is important to know that DIA data could provide enough evidence for some modifications, but may not have enough evidence to differentiate one form from the others. This is also a challenge that traditional database searching approaches have faced, with scores designed for this purpose, such as the A-score for site localization⁸. Thus, while it is possible to query for variable modifications with current implementation of PECAN, users are strongly urged to further scrutinize the results, especially if the goal is site-localization of modified forms.

Detection of modified peptides from protein N-acetylation



⁸ Beausoleil, S.A., Villén, J., Gerber, S.A., Rush, J., and Gygi, S.P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotech* 24, 1285–1292.

(a) Comparison of modified peptides of protein *N*-terminal acetylation (*N*-acetylation) detected by Comet from 4xGPF DDA data and by PECAN from 4xGPF DIA data. (b) Retention time analysis of 91 modified peptides detected by both methods.

Q5. How to interpret PECAN results when similar peptides were assign to the same group of spectra

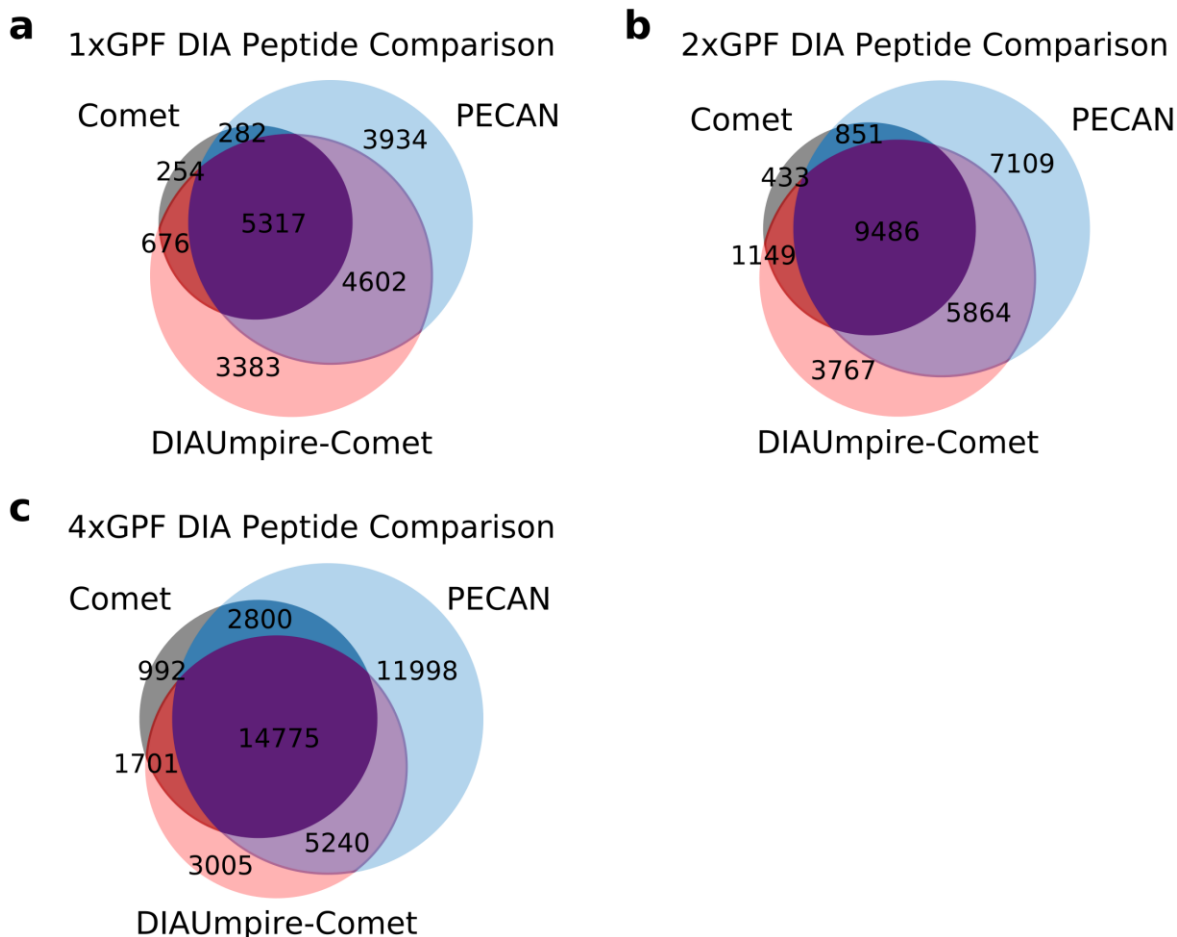
By design, PECAN assumes that the detection of one peptide is independent from the detection of other peptides. Thus, it is possible that the same group of spectra provide strong and significant evidence to more than one peptide. Depending on the isolation scheme of DIA, precursor ions of similar peptides could fall in the same isolation window. If two similar peptides share majority of their fragment ions, it is possible that they are both assigned to the same group of spectra. In such case, the result should be interpreted as both peptides are detected from the DIA data if individually they both pass the FDR (i.e. *q*-value) cutoff.

Q6. How do PECAN and DIA-Umpire workflow compare to direct Comet search allowing for wide range of precursor masses?

We analyzed the three GPF Hela DIA datasets with three methods: PECAN, Comet with wide precursor mass tolerance, and DIA-Umpire followed by Comet. For Comet analyses, precursor mass tolerance of ± 10 m/z was used for 1xGPF DIA, ± 5.0 m/z for 2xGPF DIA, and ± 2.5 m/z for 4xGPF DIA. For PECAN, precursor mass tolerance of ± 10 ppm was used for three datasets. For DIA-Umpire workflow, precursor mass tolerance of ± 10 ppm was used by the following Comet search for three datasets. When searched with the human UniProt Swiss-Prot database, Comet identified 6533, 11938, and 20276 unique peptides from 1xGPF, 2xGPF, and 4xGPF data, respectively; DIA-Umpire-Comet identified 13978, 20266, and 24721 unique peptides from 1xGPF, 2xGPF, and 4xGPF data, respectively; and PECAN detected 14135, 23398, 34813 unique peptides from 1xGPF, 2xGPF, and 4xGPF data,

respectively (shown below). In all three cases, PECAN detected more peptides compares to direct Comet search and the DIA-Umpire workflow.

Peptide comparison from Comet, PECAN, and DIA-Umpire analyses



Comparison of detected peptides by Comet, PECAN, and DIA-Umpire followed by Comet from (a) 1xGPF, (b) 2xGPF, and (c) 4xGPF DIA data when searched with the human UniProt Swiss-Prot database. For Comet analyses, precursor mass tolerance of ± 10 m/z was used for 1xGPF DIA, ± 5.0 m/z for 2xGPF DIA, and ± 2.5 m/z for 4xGPF DIA. For PECAN, precursor mass tolerance of ± 10 ppm was used for three datasets. For DIA-Umpire workflow, precursor mass tolerance of ± 10 ppm was used by the following Comet search for three datasets.

Supplementary Data

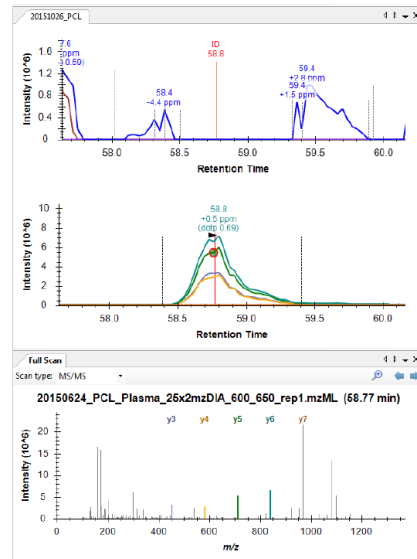
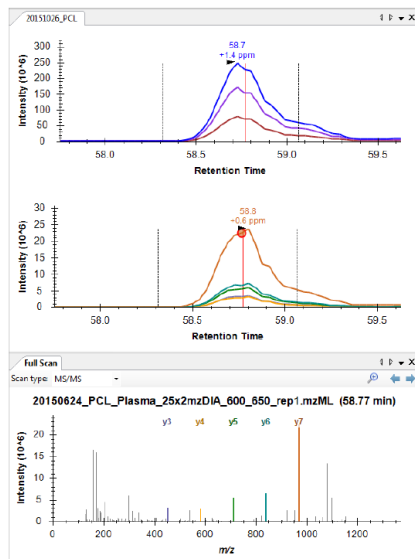
Examples of Glu to Lys variant containing peptides

a

ApoA1: E134K

Canonical peptide:
K.WQ**E**EMELYR.Q [131, 139]

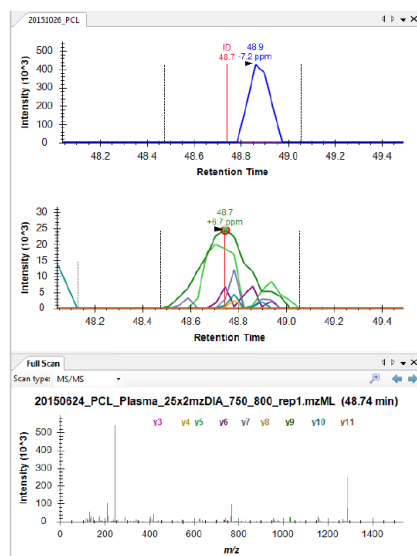
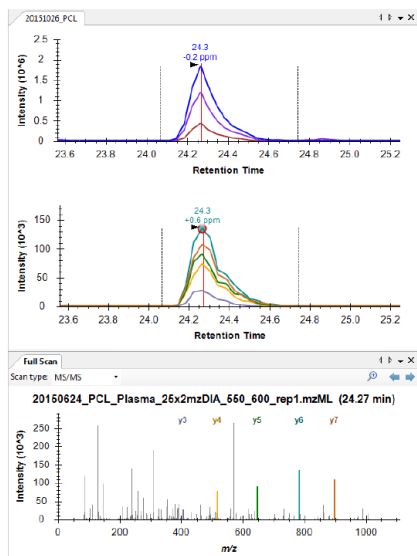
Variant peptide:
K.WQ**K**EMELYR.Q [131, 139]



b**ApoA1: E160K**

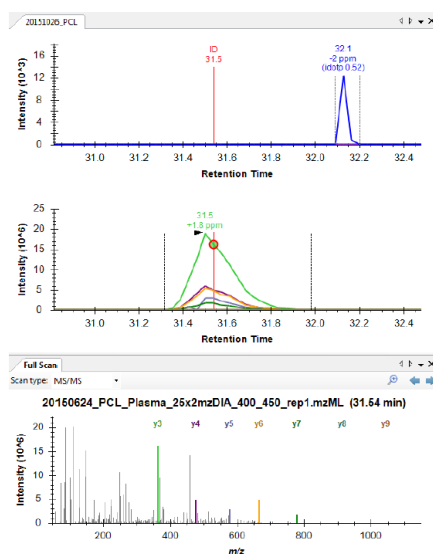
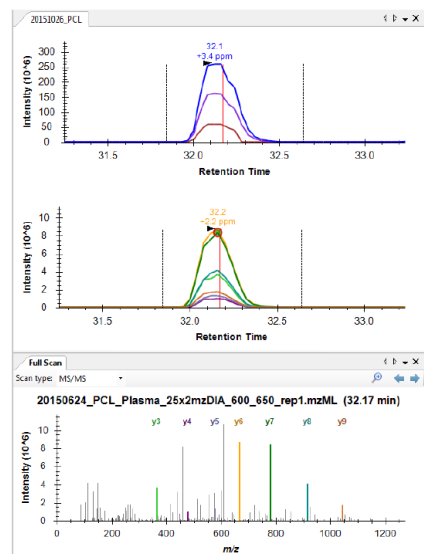
Canonical peptide:
R.QKLHELQEK.L [155, 163]

Variant peptide:
K.LQEKLSPLGEEMR.D [160, 172]

**c****ApoA1: E222K**

Canonical peptide:
K.ATEHLSTLSEK.A [219, 229]

Variant peptide:
K.ATKHLSTLSEK.A [219, 229]



Extracted ion chromatograms of precursor ions, fragment y-ions, and peptide-spectrum match (PSM) of the three variants of glutamic acid (E) to lysine (K) in Apolipoprotein A1 (ApoA1). (a) Variant peptide of E134K was detected with the same retention time, from the same group of MS/MS spectra as the canonical peptide, and shared most of the fragment ions with the canonical. The variant peptide specific y7 ion was missing from the PSM, indicating that this is likely a false positive. (b) Variant peptide of E160K was generated from the variant-specific trypsin cleavage. The fragmentation pattern of this variant peptide was distinctive from the canonical peptide. (c) Variant peptide of E222K was detected at +3 charge state whereas the canonical peptide was detected at +2 charge state.